

# Documentation

This page describes the **Prague Discourse Treebank 2.0 (PDiT 2.0)** and summarizes changes in the annotation of discourse relations carried out after the publication of the Prague Dependency Treebank 3.0 (PDT 3.0; 2013). For details on the previous versions of the Prague Dependency Treebank (PDT) and the Prague Discourse Treebank (PDiT), please refer to their respective documentations:

- PDT 2.0 documentation (<http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch05.html>)
- PDT 2.5 documentation (<http://ufal.mff.cuni.cz/pdt2.5/en/documentation.html>)
- PDiT 1.0 documentation (<http://ufal.mff.cuni.cz/pdit/documentation>)
- PDT 3.0 documentation (<http://ufal.mff.cuni.cz/pdt3.0/documentation>)

## 1 Introduction

The new version of PDiT (2.0) is enriched by the annotation of so called secondary connectives. In the PDiT 2.0, two types of discourse connectives are annotated – primary and secondary. Both of them have an ability to express semantico-pragmatic discourse relations but they differ in the degree of grammaticalization. Whereas primary connectives are grammaticalized, mostly one-word expressions (like *a* “and”, *ale* “but”, *když* “when”, *protože* “because”), secondary connectives are not (yet) fully grammaticalized expressions (cf. *z tohoto důvodu* “for this reason”, *za těchto podmínek* “under these conditions”, *kvůli tomu* “due to this” etc.). For more details, see Rysová and Rysová (2014, 2015).

Annotation of primary connectives was carried out in the first phase of our discourse annotation and was published in 2012 as PDiT 1.0 and in 2013 updated in PDT 3.0. Annotation of secondary connectives was done as a follow up of the previous versions and is included in the current version of the corpus – PDiT 2.0.

## 2 Annotation procedure

The annotation of both secondary and primary connectives share several similar features – so that it offers a suitable language material for comparison of their behaviour in authentic texts, see Rysová (2015). At the same time, due to structural differences between primary and secondary connectives, we had to establish several new aspects concerning annotation of secondary connectives that were not necessary in the first phase of annotation in PDiT 1.0.

### 2.1 Similarities in annotations of primary and secondary connectives

Similarly to primary connectives, annotation of secondary connectives includes (both inter- and intra-sentential) discourse relations that hold between two spans of a text (containing finite verbs) called discourse arguments. In the tectogrammatical level of the treebank, the relations are captured by an arrow leading between two verbal nodes (or their coordinations) representing whole arguments (see Figure 1).

Each relation is also provided by one of the semantico-pragmatic label from the classification of discourse relation established for primary connectives (like reason-result, condition, purpose etc.) and by the range of discourse arguments (i.e. the annotation includes the information about between how large spans of a text the relation holds).

### 2.2 New aspects in annotation of secondary connectives

#### a) Two-level annotation

One of the newly added aspects annotated for secondary connectives is that they are marked in two levels. The general feature of most of secondary connectives is that they contain a core word signaling given semantico-pragmatic type of discourse relation (cf. e.g. secondary connective *podmínkou je* “the condition is” containing the noun *condition* as a core word signaling a discourse relation of condition; the same core word appears also in other secondary connectives like *za těchto podmínek* “under these conditions”, *za podmínky, že* “on condition that” etc.).

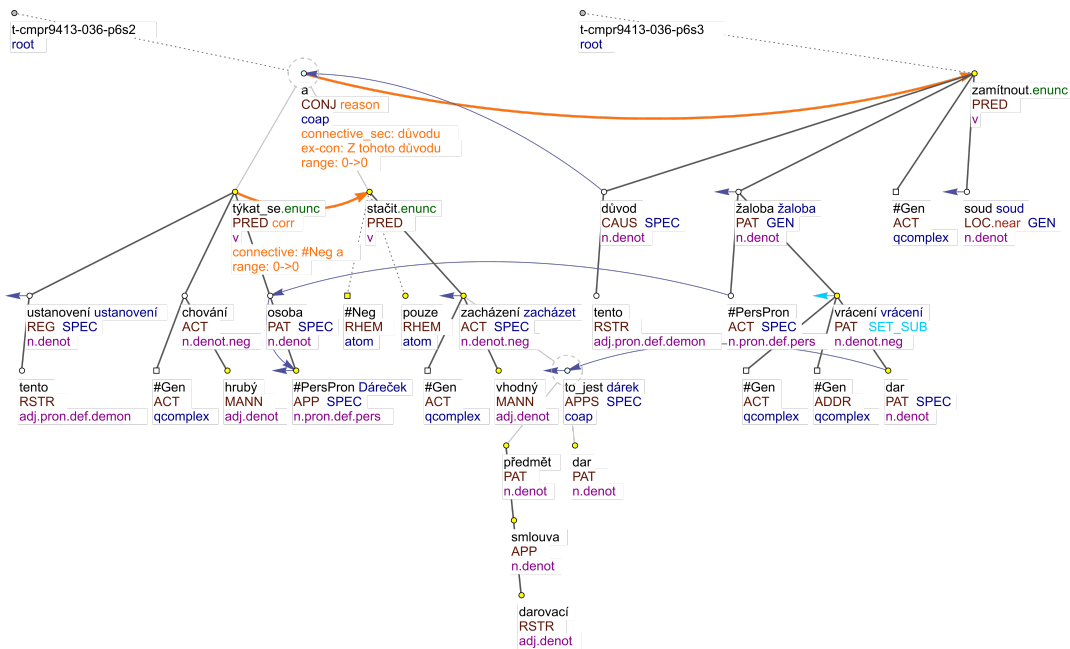
During the annotation of secondary connectives, we thus annotate firstly a suitable core word (i.e. nouns like *podmínka* “condition”, *důvod* “reason”, *výsledek* “result”, prepositions like *kvůli* “due to”, *navzdory* “besides” etc. or verbs like *znamenat* “to mean”, *shrnout* “to conclude” etc.) and then the whole structure of secondary connectives (i.e. the whole expressions like *za těchto podmínek* “under these conditions”, *z tohoto důvodu* “for this reason”, *kvůli tomu* “due to this”, *to znamená* “this means” etc.). Thanks to this annotation, it is possible to search all the possible secondary connectives containing the same core word. For illustration, see Figure 1 with a secondary connective *z tohoto důvodu* “for this reason” and core word *důvod* “reason”.

#### b) New semantic discourse relations

Also three new semantic subtypes of discourse relations were added to current annotation: relations of regard, conclusion and entailment. This indicates that secondary connectives have an ability to better express different shades of semantic relations than primary connectives.

#### c) Revision of connective types from PDiT 1.0 to PDiT 2.0: changes of some primary connectives to secondary

During the annotation, we have also revised the annotation of primary connectives in the PDiT 1.0 concerning the division of connectives into primary and secondary. Some connectives that were in PDiT 1.0 marked as primary were changed into secondary in PDiT 2.0. These expressions were non-grammaticalized structures like prepositional phrases containing demonstrative pronouns (cf. *kvůli tomu* “due to this”, *kromě toho* “besides this” etc.) or fixed phrases like *obecně řečeno* “generally speaking”, *krátce řečeno* “in short” etc.



**Figure 1:** S ohledem na toto ustanovení by se hrubé chování muselo týkat vaší osoby a nestačí pouze nevhodné zacházení s předmětem darovací smlouvy, to je darem. Z tohoto důvodu by byla vaše žaloba na vrácení daru u soudu zamítnuta.

Translation into English: With regard to this provision, the abusive behaviour would have to be related to your person and an inappropriate treatment of the subject of the donation contract is not enough. **For this reason**, your action on the return of the donation would be rejected at the court.

### 3 List of discourse-related annotation attributes in PDiT 2.0

Discourse-related annotation is captured mostly in a structured attribute *discourse* at the start node of the relation, additional annotation is captured in attributes *discourse\_groups* and *discourse\_special*.

- **discourse/target\_node.rf** – id of the target node, or undefined if there is no target node (e.g. no hypertheme in a list structure)
- **discourse/type** – the type of an arrow, two possible values: *discourse* (discourse relation), *list* (list entry)
- **discourse/start\_range** – start range of a discourse arrow; possible values: *n* where *n* (non-negative integer) = number of trees to the right of the actual tree belonging to the argument in addition to the node and its subtree (*0* means just the node and its subtree), *group* (an arbitrary set of nodes; see below attributes *discourse/start\_group\_id* and *discourse\_groups*), *forward* (means the node with its subtree plus a non-specified number of the following trees), *backward* (means the node with its subtree plus a non-specified number of the preceeding trees)
- **discourse/target\_range** – target range of a discourse arrow; possible values above
- **discourse/start\_group\_id** – identifier of a group of nodes (positive integer) where the start\_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse\_groups*
- **discourse/target\_group\_id** – identifier of a group of nodes (positive integer) where the target\_range of the arrow is set to "group"; individual nodes belonging to the group keep the group identifier in the attribute *discourse\_groups*
- **discourse/discourse\_type** – type of discourse semantic relation, such as *cond* (textual condition)
- **discourse/is\_secondary** – set to 1 if the relation is expressed by a secondary connective
- **discourse/is\_negated** – set to 1 if the relation is expressed by a negated secondary connective
- **discourse/comment** – further specifies the discourse type for some relations expressed by secondary connectives; three possible values: *Regard*, *Conclusion*, *Entailment*.
- **discourse/t-connectors.rf** – list of ids of nodes from the tectogrammatical layer that represent the discourse connective (or the core of the secondary discourse connective)
- **discourse/a-connectors.rf** – list of ids of nodes from the analytical layer that represent the discourse connective (or the core of the secondary discourse connective)
- **discourse/t-connectors\_ext.rf** – list of ids of nodes from the tectogrammatical layer that represent the whole ("extended") secondary discourse connective
- **discourse/a-connectors\_ext.rf** – list of ids of nodes from the analytical layer that represent the whole ("extended") secondary discourse connective
- **discourse\_groups** – list of identifiers of groups the given node belongs to
- **discourse\_special** – three possible values for three special roles of the phrase represented by the node and its subtree: *heading* (replaces attribute *is\_heading* from PDiT 1.0), *metatext* and *caption*.

### References

- Rysová, M.: *Diskurzivní konektory v češtině (Od centra k periférii)*. Ph.D. thesis, Charles University in Prague, Prague, Czechia, 268 pp., Oct 2015.
- Rysová, M.; Rysová, K.: Secondary Connectives in the Prague Dependency Treebank. In Hajičová, Eva; Nivre Joakim (eds.): *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University, 2015, pp. 291–299. ISBN 978-91-637-8965-6. WWW: <http://www.aclweb.org/anthology/W/W15/W15-2132.pdf> (<http://www.aclweb.org/anthology/W/W15/W15-2132.pdf>)
- Rysová, M., Rysová, K.: The Centre and Periphery of Discourse Connectives. In Aroonmanakun, Wirete; Boonkwan, Prachya; Supnithi, Thepchai (eds.): *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC 28)*. Bangkok, Thailand: Department of Linguistics, Faculty of Arts, Chulalongkorn University, 2014, pp. 452–459. ISBN 978-616-551-887-1. WWW: <http://aclweb.org/anthology/Y14-1052>

**Malostranské náměstí 25**

118 00 Praha

Czech Republic

+420 951 554 278 (phone)

+420 257 223 293 (fax)

[ufal@ufal.mff.cuni.cz](mailto:ufal@ufal.mff.cuni.cz) (<mailto:ufal@ufal.mff.cuni.cz>)



(<https://lindat.mff.cuni.cz/repository/xmlui/>)




Find us on  
Facebook

(<https://www.facebook.com/UFALMFFUK>)

Page curated by mirovsky (/~mirovsky) | Sign in (/user/login?destination=node/1357)

Institute of Formal and Applied Linguistics © 2018

Powered by  Drupal (<http://drupal.org>)